

# How Does Crowdsourced Fact-Checking Approach Tackling Misinformation Affect Audience Engagement? Evidence from Twitter's Community Notes Program

Yingxin Zhou\*  
Georgia State University  
[yzhou46@gsu.edu](mailto:yzhou46@gsu.edu)

Jingbo Hou\*  
Santa Clara University  
[jhou3@scu.edu](mailto:jhou3@scu.edu)

Yi Gao\*  
Texas Tech University  
[yi.gao@ttu.edu](mailto:yi.gao@ttu.edu)

Pei-yu Chen\*  
Arizona State University  
[peiyu.chen@asu.edu](mailto:peiyu.chen@asu.edu)

## Abstract

*The spread of online misinformation is a significant issue, prompting the use of crowdsourced fact-checking. While its impact on audience engagement with fact-checked content is well-studied, effects on audience engagement with fact-checked authors remain underexplored. Using Twitter's Community Notes program as our research context and applying counterfactual estimation, we examine how crowdsourced fact-checking influences audience engagement with fact-checked authors. Our findings reveal that community notes received by fact-checked authors can increase audience engagement with these authors in the short term, but they decrease audience engagement in the long term. This study enhances our understanding of how crowdsourced fact-checking influences audience response.*

**Keywords:** Misinformation, Crowdsourced Fact-checking, Community Notes (Birdwatch), User-Generated Content (UGC)

## 1. Introduction

In today's digital age, social media provides unparalleled access to information (Hou et al., 2024; Sabzehzar et al., 2023), but the spread of misinformation—incorrect or misleading information—has raised significant concerns among both practitioners and researchers (Muhammed & Mathew, 2022). Misinformation could lead to various issues, such as swaying public opinions, eroding trust in institutions, and wasting the time, effort, and resources of audiences (Bennett & Livingston, 2018; Gradoń, 2020; Mostagir & Siderius, 2023). Thus, combating misinformation has become increasingly important.

Online platforms employ three primary fact-checking methods to combat misinformation: computational, expert-based, and crowdsourced approaches. The computational approach uses advanced algorithms to identify misinformation at scale (Ahmed et al., 2017; Taher et al., 2022; Zhang et al., 2019), but its effectiveness is limited by the quality of training data and the algorithms' ability to consider context (Barve et al., 2022; Su et al., 2020). The expert-based approach relies on human experts for manual identification, offering a broader contextual understanding (Hameleers, 2019) but lacking scalability (Shu et al., 2017). Crowdsourced fact-checking leverages crowd wisdom to identify misinformation, combining the advantages of both prior approaches and providing relatively high efficiency and effectiveness (Saeed et al., 2022). It is also considered less biased because it incorporates diverse perspectives (Wojcik et al., 2022). Given these advantages, platforms like Twitter and Sina Weibo increasingly adopt crowdsourced fact-checking to identify misleading content (Wei et al., 2022; Wojcik et al., 2022; Wu et al., 2019). With this growing prevalence, understanding its impact on user behavior is crucial for platform development and maintaining the integrity of the online informational environment.

Previous research on the impact of crowdsourced fact-checking from the perspective of audiences has mainly concentrated on how audiences who read fact-checked content respond to fact-checked content—the content that has undergone fact-checking (Wojcik et al., 2022). However, the impact of crowdsourced fact-checking on audiences' responses to fact-checked authors—individuals whose content is fact-checked—remains underexplored. Therefore, this study focuses on a relatively underexplored question: *How does the crowdsourced fact-checking approach affect audiences' engagement with authors?*

---

\* All authors contributed equally to this study.

Investigating audience engagement with fact-checked authors is essential because it addresses a crucial gap in our understanding of the broader consequences of crowdsourced fact-checking beyond the immediate interaction with the fact-checked content itself. Audience engagement plays an important role in shaping an author's popularity by influencing the spread of their user-generated content (UGC) within online communities. This is because recommendation systems in online communities often prioritize UGC with higher engagement, thereby increasing the visibility and popularity of its authors. Conversely, UGC with lower engagement (e.g., fewer likes and comments) tends to have reduced visibility, making it difficult for its authors to gain recognition and influence. This lack of visibility hinders the ability of authors of UGC with low audience engagement to increase their popularity and build stronger ties with others, as the algorithmically driven attention is skewed toward already popular users. Consequently, authors with low audience engagement may face marginalization, as they fail to become part of the community's core social network. Given the critical role of audience engagement in shaping authors' popularity and preventing marginalization, it is essential to explore how crowdsourced fact-checking influences audience engagement with fact-checked authors.

To answer our research question, we leverage Twitter's *Community Notes* program<sup>1</sup> as our research context. Launched in 2021, the Community Notes (formerly *Birdwatch*) program allows contributors to fact-check potentially misleading tweets and annotate them with explanatory notes. Contributors can also assess the helpfulness of others' notes. Notes rated as helpful by diverse contributors are displayed under the fact-checked tweets, providing additional context and alerting audiences to potential misinformation, while unhelpful or unrated notes are not displayed. Hereafter, tweets with community notes and their authors are referred to as '*labeled tweets*' and '*labeled authors*,' respectively. Empirically, we employ counterfactual estimators (CEs) as our main identification strategy. We use CEs because they offer a more reliable estimation of treatment effect than the canonical difference-in-differences (DID) approach (Liu et al., 2024).

We find that audiences tend to interact more with labeled authors than with unlabeled ones in the short term. This suggests an unintended consequence of community notes: community notes can increase the

visibility of labeled authors who are the source of misinformation and motivate audiences to engage with labeled authors more actively. This also implies that, in the short term, crowdsourced fact-checking is unlikely to marginalize authors who are fact-checked as an author of misinformation posts.

However, in the long term, audience engagement with labeled authors decreases. The eventual decline in audience engagement suggests that the initial burst of attention does not sustain long-term audience interest. As a result, although crowdsourced fact-checking may not immediately marginalize fact-checked authors, it may still contribute to a gradual reduction in their influence over time.

Our study advances the understanding of how crowdsourced fact-checking affects audience engagement with fact-checked authors. While existing research has explored how audiences respond to fact-checked content (Wojcik et al., 2022), it often neglects the impact on audiences' attitudes and behaviors toward the authors of fact-checked content. We contribute to the literature by investing in audience engagement from the perspective of audience engagement with these authors, providing insights into the authors' potential marginalization related to crowdsourced fact-checking.

Second, our study contributes to the expanding literature on how audiences react to misinformation detected by the crowdsourced fact-checking approach. Complementing prior studies, e.g., Wojcik et al. (2022), that found that community notes can decrease the *probability* of audience engagement with misinformation at the *individual* level, our study demonstrates that community notes, however, amplify the *volume* of audience engagement with the misinformation at the *collective* level in the short term. This nuanced contrast underscores that community notes are not without limitations, emphasizing the need for careful monitoring to mitigate any adverse consequences due to their use.

## 2. Literature Review

In today's information age, misinformation has proliferated rapidly, posing significant risks to society (Bennett & Livingston, 2018; Gradoń, 2020). Fact-checking, which verifies information to ensure its accuracy, is a common approach to combating misinformation (Papanastasiou, 2020; Pennycook et al., 2020). There are three main fact-checking approaches: the computational approach, which uses

---

<sup>1</sup> Twitter. (n.d.). *Community Notes: a collaborative way to add helpful context to posts and keep people better informed.* <https://communitynotes.x.com/guide/en/about/introduction>

algorithms to verify the content's veracity; the expert approach, which relies on experts to assess content accuracy; and the crowdsourced approach, which uses crowd wisdom to assess the veracity of the content (Hameleers, 2019; Wojcik et al., 2022).

The computational approach is highly scalable and capable of processing vast amounts of data, as evidenced by numerous studies (Ahmed et al., 2017; Girgis et al., 2018; Kaliyar et al., 2020; Ozbay & Alatas, 2020; Reis et al., 2019; Shu et al., 2017; Taher et al., 2022; Zhang et al., 2019). Nevertheless, this approach struggles to accurately identify misinformation across complex contexts (Borwankar et al., 2022). Conversely, expert fact-checking, such as PolitiFact<sup>2</sup> and Snopes,<sup>3</sup> relies on the specialized knowledge of experts and is considered more effective in detecting misinformation due to their deep understanding of the fact-checking context (Hameleers, 2019). Its efficiency, however, is markedly lower than that of the computational approach, as it is constrained by the time-intensive nature of the manual checking process (Shu et al., 2017).

Due to the limitation of the prior two approaches, more platforms (e.g., Twitter and Sina Weibo) are now adopting a third approach—crowdsourced fact-checking—which leverages crowd wisdom to detect misinformation (Kim et al., 2018; Wei et al., 2022; Wojcik et al., 2022; Wu et al., 2019). This approach has moderate efficiency, which is higher than expert fact-checking and lower than computational fact-checking (Hassan et al., 2019; Saeed et al., 2022; Zhao & Naaman, 2023). Noticeably, it is as effective in identifying misinformation as expert fact-checking (Saeed et al., 2022). Furthermore, this approach exhibits low personal bias, benefiting from the diverse viewpoints of the crowd, which contributes to a more balanced and fair evaluation process (Wojcik et al., 2022).

According to prior studies (Lamprou & Antonopoulos, 2023; X. Liu et al., 2023), audiences typically perceive the crowd—comprised of laypeople—as less credible than experts, attributing this to their lesser professional knowledge and expertise. Given this difference in perceived credibility, audience reactions to crowdsourced fact-checking may differ significantly from those to expert fact-checking. For example, Dennis et al. (2023) found that expert evaluations on news source credibility can reduce audience belief in misinformation more effectively than crowdsourced evaluations. This discrepancy in the perceived credibility regarding experts and the crowd underscores the need to further

investigate the impact of crowdsourced fact-checking on audience response.

### 3. Hypotheses development

People are naturally drawn to controversial or potentially inaccurate content (Chen & Berger, 2013; Y. Liu et al., 2023; Soroka & McAdams, 2015). As a result, the presence of a community note flagging potential misinformation heightens audiences' attention and curiosity toward the labeled tweet, prompting them to explore its content and understand the reason for its flagging.

This surge in curiosity leads to a short-term increased interaction with the authors. Specifically, the immediate attention encourages audiences to interact more with labeled tweets and authors. Warnings of misinformation heighten skepticism, prompting people to critically evaluate the content (Scharrer et al., 2022). As a warning sign of potentially misleading information in a tweet, community notes push viewers to think more critically, leading them to form and express opinions and engage in discussion (Papakyriakopoulos & Goodman, 2022). Neutral audiences, who might have ignored the tweets initially, are now motivated to think, judge, and comment on these tweets. Those agreeing with the labeled tweet after reflection may support it through liking, replying, quoting, and retweeting, while those who disagree may actively respond with opposition. In this way, community notes temporarily turn passive viewers into active participants, increasing overall interaction with the labeled authors' labeled tweets in the short term.

This short period of heightened curiosity also brings in a broader audience, as community notes draw attention from individuals who might not have been interested otherwise. This expanded audience base further drives short-term interaction with both the labeled tweet and the authors' other content.

In contrast, tweets without community notes do not generate the same level of curiosity or discussion. Therefore, community notes are likely to increase audience interaction with labeled authors in the short term. Hence, we hypothesize:

**H1:** *Community notes increase the audiences' interaction with authors in the short term.*

In the short term, community notes may attract a larger audience by sparking curiosity and drawing attention to flagged content. However, in the long term, as the initial curiosity wanes, the negative impact of these notes can begin to take hold. In particular, these notes can erode the trust of loyal audiences of the labeled

<sup>2</sup> PolitiFact. (n.d.). <http://www.politifact.com/>

<sup>3</sup> Snopes. (n.d.). <http://www.snopes.com/>

authors. Loyal followers often have a positive feeling about the authors they support. However, the presence of community notes labeling their content as potentially misleading or inaccurate can be seen as a challenge to the credibility of these authors. When individuals learn about the wrongdoings of others, they form negative judgments about them (Hans & Ermann, 1989; Uhlmann et al., 2015). Thus, when loyal audiences realize their favorite authors post misinformative content, they may grow frustrated or disappointed and diminish their trust toward the authors, thereby reducing interactions with them in the long term.

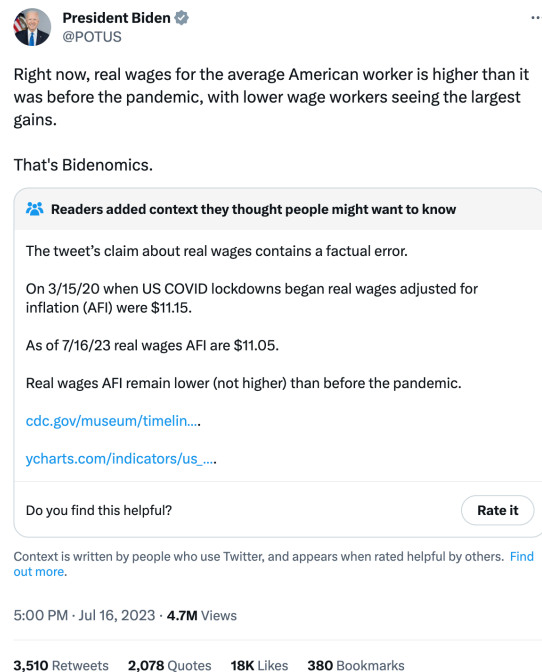
Thus, while community notes may initially increase engagement, they can harm long-term interactions by undermining trust and diminishing audience loyalty.

**H2:** *Community notes decrease the audiences' interaction with authors in the long term.*

## 4. Context, empirical strategy, and data

### 4.1. Context

On January 25, 2021, Twitter launched the *Community Notes* program, a crowdsourced fact-checking program to combat misinformation. Community Notes allows qualified users, called contributors, to identify and flag misinformation by submitting explanatory notes. Contributors can also rate the helpfulness of notes submitted by others. Notes have three statuses: 'currently rated helpful' (CRH), 'currently rated not helpful' (CRNH), and 'needs more ratings' (NMR). All proposed notes start as NMR and are hidden from public view. They are evaluated by contributors from diverse backgrounds. Once a note is widely rated as helpful or unhelpful by enough contributors, its status changes to CRH or CRNH. CRH notes are attached to the corresponding tweets, while CRNH and NMR notes are not. For example, in Figure 1, Twitter attached a community note rated CRH to President Biden's tweet was appended to his original tweet, making it visible to all Twitter users.



**Figure 1. An example of a community note**

This Community Notes program, which attaches helpful community notes to potentially misleading tweets, provides a natural experiment setting in a staggered treatment manner for our study. Notably, the implementation of the Community Notes program does not interfere with other important features of Twitter, such as content streaming (Borwankar et al., 2022), ensuring clear isolation of this Community Notes program's impact from the impacts of other system designs.

### 4.2. Identification – counterfactual estimation

To examine the impact of community notes, we compile a list of Twitter users who have received at least one community note (regardless of the note's status) between May 2022 and January 2023. We leverage the staggered rollout of community notes across labeled authors to examine their relationship with audience engagement, aligning with a staggered Difference-in-Differences (Diff-in-Diffs) framework (Burch et al., 2018).

We employ counterfactual estimators (CEs) as our identification strategy to examine the impact of the staggered rollout of community notes because they provide a more reliable estimation of the treatment effect than the traditional Diff-in-Diffs approach by relaxing certain assumptions underlying Diff-in-Diffs (Liu et al., 2024). CEs estimate the *average treatment effect on the treated* (ATT) by predicting counterfactual outcomes for treated observations

based on not-yet-treated observations. We utilize author-monthly panel data to examine the ATT of community notes on authors' audience interaction. Because of the page limit, our study focuses on the matrix-completion (MC) estimators, a type of CEs that address the issue of unobserved time-varying confounders (Kidziński & Hastie, 2023).

### 4.3. Variable definitions

Our study utilizes all tweets posted by authors from January 2022 and January 2023. This extensive sample of tweets enables a thorough investigation of the impact of community notes on audience engagement with authors.

To assess the impact of community notes on audience interaction, we utilize the audience interaction metrics with the labeled authors. These metrics include the 'likes,' 'replies,' 'retweets,' and 'quotes' received by each tweet of labeled authors, measuring the audience's engagement with tweets. We aggregate these metrics to the author-month level, calculating variables  $AudienceLike_{it}$ ,  $AudienceReply_{it}$ ,  $AudienceRetweet_{it}$ , and  $AudienceQuote_{it}$  to measure the average 'likes,' 'replies,' 'retweets,' and 'quotes' per tweet received by author  $i$  in a month  $t$ .

To account for the impact of topic distribution on audience engagement, we use the Bert model enabled by Top2Vec natural language processing package to extract topics from each tweet. This unsupervised model identifies 32,208 topics. Given the challenges of handling such a high-dimensional set of variables with traditional statistical methods, we apply a reduction function built in Top2Vec to condense the identified topics into 10 key topics. Each tweet is then assigned to one of these topics, and we calculate the average topic distribution for author  $i$  in a month  $t$ . Specifically,  $TopicDist_{itk}$  measures the percentage of tweets by author  $i$  in a month  $t$  is in the topic  $k$  out of 10 topics.

## 5. Results

Table 1 shows the results of MC estimators examining the dynamic impact of community notes on audiences' engagement with labeled authors over time. The coefficient of *1<sup>st</sup> month since Community Notes* represents the ATT of receiving community notes on the number of audience likes received by authors in the first month after authors receive their first community note. The positive and statistically

significant coefficient suggests that community notes increase the audience interaction with authors in the short term, supporting H1.<sup>4</sup> In addition, the negative and statistically significant coefficient for periods since the second month after receiving notes suggests that community notes decrease the audience interaction with authors in the long term, supporting H2.

Our MC estimator passed both the placebo test and the equivalence test, as recommended by Liu et al. (2024). These tests confirm the robustness of our estimator by ensuring that it does not produce spurious treatment effects in the absence of an actual intervention (placebo test) and does not violate the parallel trend assumptions (equivalence test). This strengthens the validity of our causal inferences.

Due to the page limit, we only report the impacts on the number of audience likes. The short-term positive and long-term negative impacts are consistent across different measures of audience engagement mentioned in Section 4.3. Figure 2 shows the dynamic ATTs in Table 1 visually.

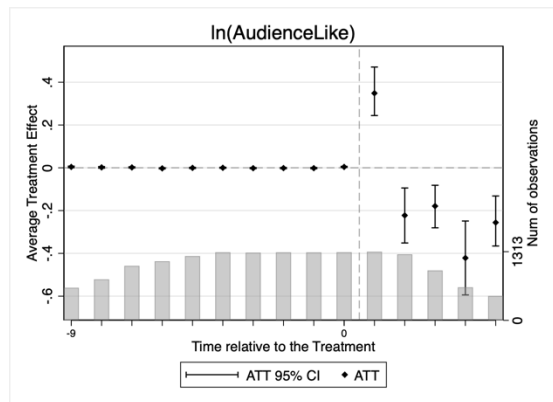
**Table 1. The dynamic effect of community notes on audience interaction**

Dep. Var.	$\ln(AudienceLike)$
<i>1<sup>st</sup> month since receiving notes</i>	0.349*** (0.061)
<i>2<sup>nd</sup> month since receiving notes</i>	-0.223*** (0.067)
<i>3<sup>rd</sup> month since receiving notes</i>	-0.179*** (0.050)
<i>4<sup>th</sup> month since receiving notes</i>	-0.422*** (0.087)
<i>5<sup>th</sup> month since receiving notes</i>	-0.256*** (0.062)
Twitter User FEs	Yes
Year-Month FEs	Yes
Placebo Test	Passed
Equivalence Test	Passed
Control for Topic Distribution	Yes
$(TopicDist_{itk})$	

Notes: Standard errors are calculated 1000 times of bootstrap runs.  
\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .

<sup>4</sup> Consistent with prior studies (e.g., Wojcik et al. 2022), we also analyze audience engagement from the tweet-reply level and find that while community notes increase the collective engagement of

the audience, they also reduce the likelihood that the audience will debunk the labeled tweets as well.



**Figure 2. The dynamic effect of community notes on audience interaction**

## 6. Conclusion

Our study has an interesting finding with important implications. While community notes can decrease audience engagement with labeled authors over the long term, we find that these notes can actually boost engagement with authors creating misinformation in the short term. Thus, while community notes serve as a tool to combat misinformation, they may also unintentionally heighten the visibility and interaction with the very authors they aim to fact-check in the short run. Addressing this short-term surge in audience engagement is crucial for developing more effective strategies to mitigate the spread of misinformation.

This study contributes to the literature by investigating the impact of the crowdsourced fact-checking approach from the perspective of audience engagement with authors. The existing literature on the crowdsourced fact-checking approach has primarily focused on its impact on labeled tweets (Wojcik et al., 2022). We extend this body of work by examining the impact of community notes on audience engagement with labeled authors.

Our study has important managerial implications for social media platforms. First, social media platforms should be aware of the potential unintended effects of community notes on audience interaction. We observe that audiences engage more with an author's posts in the short term right after the author receives community notes, and this indicates that crowdsourced fact-checking can inadvertently increase audience engagement with labeled authors. Because recommendation systems used by social media platforms often prioritize UGC with high audience engagement, this unintended surge in audience engagement could unintentionally amplify the spread of misinformative UGC. Social media platforms should consider this unintended boost in

audience engagement when developing their recommendation systems.

Second, the marginalization issue of labeled authors due to the adoption of crowdsourced fact-checking could be a challenge that social media platforms would face. Because recommendation systems often prioritize UGC with higher engagement, UGC with lower engagement receives less attention, limiting its authors' recognition and influence. This may marginalize authors with low engagement. Our findings show that community notes decrease engagement with labeled authors in the long run, suggesting that their potential marginalization due to decreased audience interaction could be a potential concern for platforms considering the adoption of crowdsourced fact-checking.

## References

- Ahmed, H., Traore, I., & Saad, S. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*.
- Barve, Y., Saini, J. R., Kotecha, K., & Gaikwad, H. (2022). Detecting and Fact-Checking Misinformation Using "Veracity Scanning Model". *International Journal of Advanced Computer Science and Applications*, 13(2).
- Bennett, W. L., & Livingston, S. (2018). The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions. *European journal of communication*, 33(2), 122-139.
- Borwankar, S., Zheng, J., & Kannan, K. N. (2022). Democratization of Misinformation Monitoring: The Impact of Twitter's Birdwatch Program. *Available at SSRN 4236756*.
- Burtch, G., Carnahan, S., & Greenwood, B. N. (2018). Can You Gig It? An Empirical Examination of the Gig Economy and Entrepreneurial Activity. *Management science*, 64(12), 5497-5520.
- Chen, Z., & Berger, J. (2013). When, Why, and How Controversy Causes Conversation. *Journal of Consumer Research*, 40(3), 580-593.
- Dennis, A. R., Moravec, P. L., & Kim, A. (2023). Search & Verify: Misinformation and Source Evaluations in Internet Search Results. *Decision Support Systems*, 171, 113976.
- Girgis, S., Amer, E., & Gadallah, M. (2018). Deep Learning Algorithms for Detecting Fake News in Online Text. 2018 13th international conference on computer engineering and systems (ICCES).
- Gradoñ, K. (2020). Crime in the Time of the Plague: Fake News Pandemic and the Challenges to Law-Enforcement and Intelligence Community. *Society Register*, 4(2), 133-148.
- Hameleers, M. (2019). Susceptibility to Mis-and Disinformation and the Effectiveness of Fact-Checkers:

- Can Misinformation Be Effectively Combated? *SCM Studies in Communication and Media*, 8(4), 523-546.
- Hans, V. P., & Ermann, M. D. (1989). Responses to Corporate Versus Individual Wrongdoing. *Law and Human Behavior*, 13(2), 151-166.
- Hassan, N., Yousuf, M., Mahfuzul Haque, M., A. Suarez Rivas, J., & Khadimul Islam, M. (2019). Examining the Roles of Automation, Crowds and Professionals Towards Sustainable Fact-Checking. Companion Proceedings of The 2019 World Wide Web Conference.
- Hou, J., Liang, C., & Chen, P.-y. (2024). How Socially Perceived Threat Shapes Preventive Behavior in the Context of Covid-19. *Production and Operations Management*, 10591478241231864.
- Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). Fndnet—a Deep Convolutional Neural Network for Fake News Detection. *Cognitive Systems Research*, 61, 32-44.
- Kidziński, Ł., & Hastie, T. (2023). Modeling Longitudinal Data Using Matrix Completion. *Journal of Computational and Graphical Statistics*, 1-16.
- Kim, J., Tabibian, B., Oh, A., Schölkopf, B., & Gomez-Rodriguez, M. (2018). Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation. Proceedings of the eleventh ACM international conference on web search and data mining.
- Lamprou, E., & Antonopoulos, N. (2023). Ranked by Truth Metrics: A New Communication Method Approach, on Crowd-Sourced Fact-Checking Platforms for Journalistic and Social Media Content. *Stud Media Commun*, 11, 231-243.
- Liu, L., Wang, Y., & Xu, Y. (2024). A Practical Guide to Counterfactual Estimators for Causal Inference with Time - Series Cross - Sectional Data. *American Journal of Political Science*, 68(1), 160-176.
- Liu, X., Qi, L., Wang, L., & Metzger, M. J. (2023). Checking the Fact-Checkers: The Role of Source Type, Perceived Credibility, and Individual Differences in Fact-Checking Effectiveness. *Communication Research*, 00936502231206419.
- Liu, Y., Zhang, J., & Zhang, P. (2023). Negativity Bias During Information Seeking, Processing, and Sensemaking About a Policy Debate: An Eye - Tracking Experiment. *Proceedings of the Association for Information Science and Technology*, 60(1), 267-278.
- Mostagir, M., & Siderius, J. (2023). Social Inequality and the Spread of Misinformation. *Management Science*, 69(2), 968-995.
- Muhammed , T. S., & Mathew, S. K. (2022). The Disaster of Misinformation: A Review of Research in Social Media. *International journal of data science and analytics*, 13(4), 271-285.
- Ozbay, F. A., & Alatas, B. (2020). Fake News Detection within Online Social Media Using Supervised Artificial Intelligence Algorithms. *Physica A: statistical mechanics and its applications*, 540, 123174.
- Papakyriakopoulos, O., & Goodman, E. (2022). The Impact of Twitter Labels on Misinformation Spread and User Engagement: Lessons from Trump's Election Tweets. Proceedings of the ACM web conference 2022,
- Papanastasiou, Y. (2020). Fake News Propagation and Detection: A Sequential Model. *Management Science*, 66(5), 1826-1846.
- Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines without Warnings. *Management science*, 66(11), 4944-4957.
- Reis, J. C., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised Learning for Fake News Detection. *IEEE Intelligent Systems*, 34(2), 76-81.
- Sabzehzar, A., Zhou, Y., & Hou, J. (2023). Can Social Disconnectedness Inhibit Online Trade? Examining the Effects of Digital Distance on Peer-to-Peer Lending.
- Saeed, M., Traub, N., Nicolas, M., Demartini, G., & Papotti, P. (2022). Crowdsourced Fact-Checking at Twitter: How Does the Crowd Compare with Experts? Proceedings of the 31st ACM International Conference on Information & Knowledge Management,
- Scharrer, L., Pape, V., & Stadler, M. (2022). Watch Out: Fake! How Warning Labels Affect Laypeople's Evaluation of Simplified Scientific Misinformation. *Discourse Processes*, 59(8), 575-590.
- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD explorations newsletter*, 19(1), 22-36.
- Soroka, S., & McAdams, S. (2015). News, Politics, and Negativity. *Political communication*, 32(1), 1-22.
- Su, Q., Wan, M., Liu, X., & Huang, C.-R. (2020). Motivations, Methods and Metrics of Misinformation Detection: An Nlp Perspective. *Natural Language Processing Research*, 1(1-2), 1-13.
- Taher, Y., Moussaoui, A., & Moussaoui, F. (2022). Automatic Fake News Detection Based on Deep Learning, Fasttext and News Title. *International Journal of Advanced Computer Science and Applications*, 13(1).
- Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A Person-Centered Approach to Moral Judgment. *Perspectives on Psychological Science*, 10(1), 72-81.
- Wei, X., Zhang, Z., Zhang, M., Chen, W., & Zeng, D. D. (2022). Combining Crowd and Machine Intelligence to Detect False News on Social Media. *Mis Quarterly*.
- Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M., Coleman, K., & Baxter, J. (2022). Birdwatch: Crowd Wisdom and Bridging Algorithms Can Inform Understanding and Reduce the Spread of Misinformation. *arXiv preprint arXiv:221015723*.
- Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in Social Media: Definition, Manipulation, and Detection. *ACM SIGKDD explorations newsletter*, 21(2), 80-90.
- Zhang, C., Gupta, A., Kauten, C., Deokar, A. V., & Qin, X. (2019). Detecting Fake News for Reducing Misinformation Risks Using Analytics Approaches. *European Journal of Operational Research*, 279(3), 1036-1052.
- Zhao, A., & Naaman, M. (2023). Insights from a Comparative Study on the Variety, Velocity, Veracity, and Viability of Crowdsourced and Professional Fact-

Checking Services. *Journal of Online Trust and Safety*,  
2(1).